An Advanced Hidden Markov Model for Hourly Rainfall Time Series

Oliver Stoner[1] and Theo Economou

February 3, 2020

Department of Mathematics, University of Exeter, UK.
[1]O.R.Stoner@exeter.ac.uk

Abstract

The hidden Markov framework is adapted to construct a compelling model for simulation of sub-daily rainfall, capable of capturing important characteristics of sub-daily rainfall well, including: long dry periods or droughts; seasonal and temporal variation in occurrence and intensity; and propensity for extreme values. These adaptations include both clone states and temporally non-homogeneous state persistence probabilities. Set in the Bayesian framework, a rich quantification of parametric and predictive uncertainty is available, and thorough model checking is made possible through posterior predictive analyses. Results from the model are highly interpretable, allowing for meaningful examination of diurnal, seasonal and annual variation in sub-daily rainfall occurrence and intensity. To demonstrate the effectiveness of this approach, both in terms of model fit and interpretability, the model is applied to an 8-year long time series of hourly observations.

Keywords: Extreme values; droughts; non-homogeneous; persistence; simulation; sub-daily.

1 Introduction

Severe flooding events, such as those that occurred in the UK in the winter of 2013-2014 and the winter of 2015-2016, pose a great risk to society. For each winter, the total economic damage caused by the flooding was estimated to be over one billion pounds (GBP) (Environment Agency (2016), Environment Agency (2018)). Hydrological flood models play an important role in mitigating this damage, for example by helping to inform the planning of new flood defences and drainage systems, as well as integration with flood warning systems.

Typically, hydrological models are used to test the response of the hydrological system to design-storms, which are intended to represent an idealised extreme rainfall scenario. However, Chandler et al. (2014) argue that this approach is limited, in the first instance because the temporal profile of the design-storm may fail to capture important characteristics of system performance. Moreover, focusing on the response of the system to a single event may be inadequate, as the risk of flooding posed by a single storm event depends strongly on the antecedent conditions of the catchment (Chandler et al., 2014). For example, the risk of flooding may depend on whether or not the catchment has already been saturated by recent rainfall. For this reason, Chandler et al. (2014) argue that hydrological models should instead use long rainfall time series generated from stochastic/probabilistic models as inputs, so that the effects of both the rainfall intensity during a storm event and antecedent conditions can be taken into account. In particular, Segond et al. (2007) argue that while daily rainfall time series may be suitable as inputs for flood modelling in rural catchments, for urban catchments a higher temporal resolution is necessary, as the response of the hydrological system may develop on a shorter time-scale. This motivates the development of rainfall time series models at sub-daily resolutions, e.g. hourly.

Modelling rainfall is challenging because its natural variability can dominate any seasonal structures or long-term temporal trends, more so than other meteorological variables such as temperature and wind (Chandler et al., 2014). For this reason, it is vitally important to capture well the whole distribution of rainfall values, which is in itself a non-trivial task due to the propensity of extremely high rainfall values. This is illustrated in the left plot of Figure 1, which shows empirical quantiles of non-zero rainfall gauge observations from an 8-year hourly time series in Exeter, UK (Met Office, 2019). The heavy tail is especially clear when comparing the highest hourly values (well over 20mm) to the 99% quantile, which is only 5.2mm. Ensuring the model is capable of reproducing these extremes is essential in any application where they are of great concern, such as surface flood modelling. An equivalent return-level plot can be found in the supplementary material.

A further modelling challenge is the complex temporal structure in the occurrence of

rainfall, which notably consists of long dry periods or droughts. To illustrate this, the right plot in Figure 1 shows the empirical quantiles of dry period lengths (defined as periods where the hourly rainfall value does not exceed 0.2mm). This distribution also has quite a heavy tail, including several high values which may be considered extremes. Capturing these extremes is equally important, particularly for applications where droughts are a concern, such as agricultural planning.

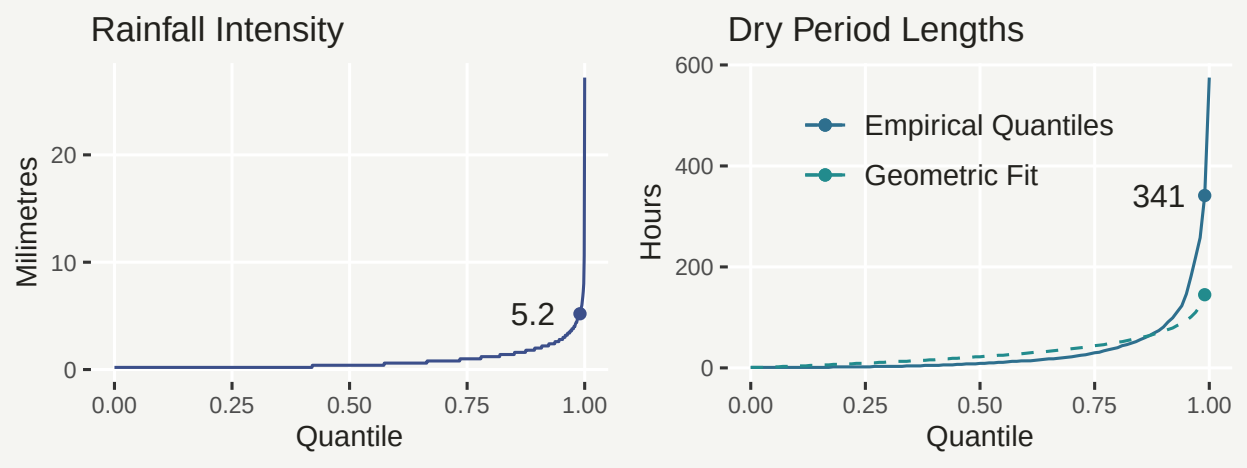

Figure 1: **Left:** Empirical quantiles of the hourly rainfall intensities (observations strictly greater than 0mm) from the Exeter gauge time series. **Right:** Empirical quantiles of dry period lengths (for this plot taken to be rainfall not exceeding 0.2mm in any given hour), and theoretical quantiles of a Geometric fit (dashed line). **Points:** 99% empirical quantiles.

Finally, both rainfall intensity and occurrence can vary substantially by season. In the UK for instance, summer rainfall extremes tend to be more severe than in the winter (Chandler et al., 2014). For this reason, care should be taken to ensure model simulations accurately reflect the time of year.

In this article we propose a flexible model, based on advanced hidden Markov latent states, which is capable of capturing key features of sub-daily rainfall well:

- Diurnal (time of day), seasonal, and temporal variation in occurrence and intensity;
- Long dry periods (which vary with time and season);
- Extreme values (which vary with time and season).

Given that such stochastic rainfall models are often used in decision making (e.g. warnings) or as inputs to physical models (e.g. hydrological), a further requirement is that parametric uncertainty is fully quantified and propagated into the simulations. To achieve this we

implement the proposed model in the Bayesian hierarchical framework. Applied to 8 years of hourly rainfall gauge data, we demonstrate using posterior predictive model checking how our proposed model performs well overall and specifically with respect to capturing these three features.

The article is structured as follows: Section 2 gives an overview of existing approaches to stochastic/statistical modelling of rainfall, highlighting their strengths and limitations. In Section 3, we propose a flexible model for sub-daily rainfall gauge time series, which we then apply to an 8-year long time series of hourly observations. In Section 4, we present extensive model checking and an analysis of model output. Finally, the article ends with a critical discussion of our approach in Section 5.

2 Background

In the previous section we outlined features of rainfall occurrence and intensity, namely long dry periods and extremely high values, which make modelling challenging. The literature on statistical modelling of rainfall is vast, and comprehensive reviews are available in Chandler et al. (2014) and Wilks and Wilby (1999). Here we seek only to give a brief overview to establish our contribution. To this end, we focus on approaches which take into account the temporal structure of rainfall and those which model both occurrence and intensity. Most of these can be roughly separated into two classes: direct and indirect rainfall models.

2.1 Direct rainfall models

The first way of modelling rainfall time series is to characterise the occurrence and amount of rainfall at a given time as random quantities, arising directly from some probabilistic framework – typically some form of regression.

One such approach (e.g. Yang et al. (2005)) consists of two Generalized Linear Models (GLM) (Dobson and Barnett, 2018), one for occurrence and one for intensity. First the data (typically daily rainfall) is transformed into a sequence of 0s (zero rainfall) and 1s (greater than zero rainfall). The model for occurrence is then a logistic regression for the probability of rain p_i on time step i:

$$\log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x}_i'\boldsymbol{\beta}. \tag{1}$$

where $\boldsymbol{\beta}$ is a vector of coefficients corresponding to covariates $\boldsymbol{x}_i$. Then, conditional on the occurrence of rain, the mean amount of rain for wet day j is modelled by a Gamma

regression:

$$\log(\lambda_j) = \boldsymbol{\xi}_j' \boldsymbol{\gamma}. \tag{2}$$

where $\boldsymbol{\gamma}$ is a vector of coefficients corresponding to covariates $\boldsymbol{\xi}_j$. The shape parameter is assumed constant.

Covariates $\boldsymbol{x}_i$ and $\boldsymbol{\xi}_j$ can include climatological variables such as the North Atlantic Oscillation (NAO), and seasonal variation can be captured through covariates such as the month. Temporal structure can be introduced in the model for occurrence by including in $\boldsymbol{x}_i$ binary variables representing the occurrence of rain in the previous k days, equivalent to a k-order Markov model (Yang et al., 2005). Similarly, $\boldsymbol{\xi}_j$ can include the rainfall amount from previous days to induce temporal structure in rainfall intensity.

An advantage of this approach is that, set in the GLM framework, models can be implemented quickly, both in terms of the computational cost of fitting and the relatively little coding required to construct and modify them. Furthermore, these models have been extended to incorporate spatial structures, for applications to data from multiple sites (Yang et al., 2005). Further extensions involve temporally structured random effects (Glasbey and Nevison, 1997) and distributions outside the exponential family (Serinaldi and Kilsby, 2012).

However, the main downside of this approach is that the models for occurrence and intensity are separated. This may be a valid approach for coarser temporal resolutions (e.g. daily rainfall), but for sub-daily rainfall it may result in failure to capture important patterns. For example, following a period of no rainfall, the arrival of a storm might lead to initially low rainfall intensity, which increases over a few hours before reaching a maximum and then decreasing. Being able to simulate such patterns may be essential for testing the response of a drainage system to a severe storm, for instance.

2.2 Indirect rainfall models

An alternative approach is to instead stochastically model the larger structures which generate rainfall, such as storm events. Given the occurrence of such structures, the amount of rainfall over a time period is then deterministic, based on some simplifying assumptions (see for instance Rodriguez-Iturbe et al. (1987)). Several parameters control different aspects of the model, such as storm event duration, frequency of occurrence, and intensity. Through these parameters, it is possible to tune the model to capture certain properties of the observed data, such as the length of dry periods, temporal dependence and the distribution of non-zero rainfall.

Whilst these approaches consist of few parameters and are in some sense based on physical justification, the absence of a likelihood function on which to base inference makes imple-

mentation challenging. Thus it is often necessary to rely on alternative methods of implementation, such as by maximising objective functions (potentially missing out on parametric uncertainty) or by employing Approximate Bayesian Computation (ABC), which requires the careful selection of summary statistics (Aryal and Jones, 2017). Furthermore, it is not evident that these approaches are able to reproduce extreme rainfall values well, a feature which Chandler et al. (2014) argue is lacking in the sub-daily rainfall modelling literature.

For comprehensively capturing multiple aspects of rainfall time series, we argue that direct characterisation of the distribution offers a more flexible framework than indirect modelling, for example by allowing different aspects of rainfall (e.g. the persistence of dry periods) to be characterised as a function of various temporal structures, including seasonality. In addition, model implementation, expansion, and checking are much more straightforward, as illustrated in Section 4. For these reasons, we pursue a fully probabilistic framework to directly model the distribution of rainfall in time.

2.3 Hidden Markov models (HMMs)

A further family of direct probabilistic models that has been used extensively for rainfall data (e.g. Kim and Lee (2017), Rayner et al. (2016)), to capture temporal structures in both rainfall occurrence and intensity, is hidden Markov models (HMMs). In HMMs, a hidden, unobservable quantity z_t varies over discrete time steps, alternating between a finite number Z of values or states $z_t \in \{S_1, ...S_Z\}$. Variable z_t is a discrete Markov chain, whose evolution over time is probabilistic, governed by a transition matrix P of probabilities. The particular state the hidden variable is in at a given time step affects (the parameters of) the conditional model for the observed quantity, which in the case of rainfall translates to the model(s) for occurrence and intensity.

HMMs are useful for rainfall because they can capture its temporal behaviour through the Markovian structure of the latent chain, without the need to explicitly include climatological structures, such as the arrival of weather fronts, or other physical processes. However in the same vain as indirect rainfall models they offer a high degree of interpretability, in that the latent states can represent weather features such as dry periods or different stages in the arrival of storms. This is aided by the fact that occurrence and intensity are both driven by the same latent states.

Conventional HMMs for rainfall often suffer from a number of shortcomings, such as underestimation of the length of long dry periods (Chandler et al., 2014). However, their flexibility as a framework, afforded by the freedom to specify virtually any conditional model for occurrence and intensity, means that it may be possible to address these issues through

a number of extensions. Based on this idea, in the subsequent section we will present our approach to modelling rainfall gauge data, which we argue is capable of capturing all the key features of sub-daily rainfall time series identified in Section 1.

3 Methodology

It is instructive to begin by defining a basic and generic HMM for rainfall. Capturing the whole distribution (tails and bulk) of rainfall well is challenging, but one way of doing so is through a discrete mixture of, say $Z = 3$, distributions. These can be interpreted as rainfall severity states, i.e. "dry", "wet", "wetter". Here "dry" is defined as a period of zero or very little rainfall, and "wet" is defined as a period of predominantly non-zero rainfall. A discrete random quantity $z_t \in \{1, 2, 3\}$ is used to characterize the distribution of rainfall x_t as:

$$p(x_t) = \sum_{j=1}^{Z} \mathbb{1}(z_t = j) p(x_t|z_t). \tag{3}$$

where $\mathbb{1}$ is the identity function, and $p(x_t|z_t)$ is the conditional distribution of rainfall x_t for each state. HMMs allow for temporal dependence by assuming that z_t is an unobserved discrete Markov chain, so that temporal structure is introduced in the persistence of each state. This is parametrised by a transition matrix $P = \{p_{i,j}\}$ where $p_{i,j} = \Pr(z_t = j|z_{t-1} = i)$.

Such models are conventionally homogeneous, meaning that the transition between states in the HMM is time invariant. However, this does not allow for the effect of seasonal variation or climatological covariates on the temporal structure of rainfall. Several articles (such as Spezia (2006), Meligkotsidou and Dellaportas (2011) and Antonello and Roberto (2012)) have instead presented non-homogeneous hidden Markov models (NHMMs), where covariates are used to characterise the parameters of the transition matrix. This added flexibility could be used to allow for seasonal or long term heterogeneity in the temporal structure of rainfall.

Additionally, HMMs are restricted by the fact that the number of time steps that the hidden variable z_t persists in a given state is implicitly Geometrically distributed. The right plot of Figure 1 illustrates that Exeter's dry period length distribution has a very heavy tail, with several dry periods lasting hundreds of hours. This would be a concern if the HMM consists of only one dry state and hence relies on an implicit Geometric model to capture this distribution, which we illustrate by also plotting a method of moments Geometric fit to the dry period lengths.

Including additional unique 'dry states', which the hidden state parameter could transition between, may introduce sufficient flexibility to capture the longest dry periods. However,

this would impede the physical interpretability of the model and potentially introduce identifiability issues. In the case that the dry period distribution is seasonally varying, then some improvement may be possible by introducing non-homogeneity in the dry state, though this may still be insufficient.

A more potent solution would instead be to use a hidden semi-Markov model (HSMM), where the persistence distribution is explicitly defined and thus can be chosen to have a heavier tail. However, HSMMs are often impractical and too computationally expensive to implement. This is especially true when the total number of time steps T is large, as many of the implementation methods have an algorithmic order $O(Z^2T^2)$ or even $O(Z^2T^3)$, compared to only order $O(Z^2T)$ for HMMs. Often, an upper bound is imposed on the persistence time distribution prior to fitting the model, to ensure computational feasibility (Economou et al., 2014). However, this strong prior statement about the persistence distribution can lead to invalid parameter estimates (Dewar et al., 2012). While this issue can be overcome by making the restriction adaptive in the implementation process, the method is still of far greater computational complexity than the basic HMM.

Once an appropriate choice of temporal structure is made, it remains to specify the conditional rainfall model. This usually involves the mixture of a Bernoulli quantity, representing the occurrence of rainfall, and a strictly positive quantity, representing the intensity of rainfall, conditional on occurrence. The choice of distribution for the intensity is made difficult by the presence of extremely high observations. For many uses of output from a statistical rainfall model, including urban flood modelling, the risk posed by these extreme events are of particular concern. Furrer and Katz (2008) argue that the commonly used Gamma distribution is not able to capture these extremes well. Other examples include the Weibull distribution (Bruno et al., 2008), although Furrer and Katz (2008) concluded that this distribution is also inadequate.

Several approaches, such as Li et al. (2012) aim to better capture extremes by mixing a more typical distribution, in this case the Exponential distribution, with the Generalised Pareto distribution for values above a given threshold, which is estimated by imposing a continuity constraint on the two distributions. However, Furrer and Katz (2008) also found that these approaches, while still performing better than the Gamma and Weibull distributions, are not able to capture well the likelihood of extreme values.

In what follows, we present an extended HMM-based framework that is flexible enough to adequately capture both temporal persistence and extremes in rainfall while retaining interpretability.

3.1 Clone states and non-homogeneity

We begin by considering the basic three-state HMM introduced earlier. One state is intended to capture dry periods, while the remaining two are intended to capture wet periods. These wet states may end up representing periods of low and high rainfall intensity, respectively, or may differ in how long they last or how often they occur. The HMM structure for z_t is defined by an initial state probability vector P_0 and a transition matrix P:

$$P = \begin{pmatrix} p & q_1(1-p) & q_2(1-p) \\ r_{1,0} & r_{1,1} & r_{1,2} \\ r_{2,0} & r_{2,1} & r_{2,2} \end{pmatrix}. \tag{4}$$

For reasons that will soon be clear, we parametrise the first row (corresponding to the dry state) in terms of the probability of remaining in the dry state (p), and the conditional (on a transition out of the dry state) probabilities of transitioning into each wet state, q_1 and q_2 such that $q_1 + q_2 = 1$. As discussed in Section 2, a restriction of this model is that the length of time spent in the dry state has an implicit Geometric(p) distribution, which may not be sufficiently flexible to capture long dry periods. For a long time series, such as $t = 1, \ldots, 70128$ hours in our application later on, an HSMM framework is prohibitively computationally expensive, so here we look for ways to retain the practicality of HMMs while making them more flexible in terms of capturing long persistence periods.

Zucchini et al. (2017) present a way of achieving a more flexible persistence distribution for a given state, without losing the convenience of the HMM framework, which we discuss here in the context of the basic HMM for rainfall. For clarity of exposition, denote the dry state as d and the wet states as w_j, for $j = 1, 2$. The idea is to introduce a number of "clone" dry states $d_1, \ldots, d_D$, which are all identical to each other and to the original dry state d in the sense that they all have the same conditional model. The transition matrix is then defined in such a way that transitions from w_j are only possible to the first clone state d_1. From here, the hidden chain can persist in the first clone state d_1 with probability p_1, or transition to w_j, or transition to the second clone state d_2. If the chain transitions to d_2, it can remain there with probability p_2, or transition to w_j, or to the next clone state d_3, and so on. The motivation behind this approach is that, while the number of time steps spent in each of the clone states is still Geometric, the total amount of time spent in any of the clone states before transitioning to another unique state w_j is a more flexible distribution – essentially a weighted sum of Geometric distributions. Note that the dry state d is now only implicitly defined in the sense that all clone states relate to the same conditional model for "dryness" (low or zero rainfall).

Here we also employ an approach that is based on the introduction of clone states. However, our formulation is such that it allows for modelling flexibility, particularly in terms of introducing temporal non-stationarity in the persistence of the dry state. To that end we consider the following constrained transition matrix:

$$P = \begin{pmatrix} p_1 & 0 & \dots & 0 & q_1(1-p_1) & q_2(1-p_1) \\ 0 & p_2 & \dots & 0 & q_1(1-p_2) & q_2(1-p_2) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & p_D & q_1(1-p_D) & q_2(1-p_D) \\ v_1 r_{1,0} & v_2 r_{1,0} & \dots & v_D r_{1,0} & r_{1,1} & r_{1,2} \\ v_1 r_{2,0} & v_2 r_{2,0} & \dots & v_D r_{2,0} & r_{2,1} & r_{2,2} \end{pmatrix}. \tag{5}$$

Here, transitions are possible from wet states w_j into any of the clone dry states d_i, while no transitions between the clone states are possible. The latter is achieved by constraining the off-diagonal entries of the first D rows and columns to be zero. As such, while the dry state persistence distributions are each Geometric(p), parameter p can now be thought of as a (categorical) random quantity taking values in $\{p_1, p_2, \dots, p_D\}$ such that the marginal distribution for the time spent in the implicit dry state is, like the approach in Zucchini et al. (2017), a more flexible Geometric mixture.

To ensure that it is possible to interpret the clone dry states as a single state, further constraints are imposed on the transition matrix: First, conditional on a transition from a dry state to a wet state, the transition probabilities (q_1 and q_2) into each wet state are invariant of the dry state. Second, conditional on a transition from a wet state to a dry state, the transition probabilities $(v_1, ..., v_D)$ into each dry state are invariant of the wet state. These are in addition to the constraint that the conditional model for rainfall occurrence and intensity is the same for all of the dry states.

This approach is equivalently flexible to the one from Zucchini et al. (2017), in the sense that it can better capture heavy tailed persistence distributions. This implies that, without sacrificing the physical interpretability of having only one dry state, or the practicality of the HMM framework, extra flexibility is afforded to potentially better capture the longest dry periods. However, as the parameters of the transition matrix are time-constant, the model can't capture seasonal or annual variation in the expected length of dry periods which may be, for example, longer on average in the summer than in the winter. The advantage of our approach is that it is straightforward to directly model the dry state persistence probabilities $p_1, ..., p_D$ as temporally-varying. One way of achieving this is a logistic model for the dry

state persistence probabilities:

$$\log\left(\frac{p_d(t)}{1-p_d(t)}\right) = u(t,d). \tag{6}$$

Here $u(t,d)$ represents a general model of time $t = 1, ..., T$ and hidden state $d = 1, ..., D$, which may also include covariate effects, including large-scale climate indices such as the North Atlantic Oscillation (NAO).

For our application to the Exeter data, we characterise $u(t,d)$ by combining an intercept term $\iota(d)$, which is different for each clone dry state, and a number of penalised splines and random effects which are common across clone states:

$$u(t,d) = \iota(d) + a_1(t) + a_2(t) + m_a(t) + y_a(t). \tag{7}$$

Function $a_1(t)$ is a cyclic (the two end points have equal value) cubic spline of the time-of-day, which is intended to capture the diurnal cycle in the dry period length. Function $a_2(t)$ is also cyclic but for the time-of-year, aimed at capturing smooth seasonal variation, while $m_a \sim \mathrm{N}(0, \delta^2_{m_a})$ are i.i.d. monthly random effects (one for each month, e.g. the 'January' effect is shared across all Januaries in the time series), which aim to capture non-smooth within-year structured variability. Lastly, $y_a \sim \mathrm{N}(0, \delta^2_{y_a})$ are i.i.d. yearly random effects aimed at capturing non-smooth between-year variability. The use of both splines and random effects affords the model flexibility to capture different (smooth and non-smooth) dry period persistence structures which may occur in different climatic conditions, but also to better capture very long dry periods.

Choosing the number of latent states is, to a certain degree, subjective. In general, more states will result in a better fit to the data, but care needs to be taken to avoid over-parametrisation and over-fitting. In general, we advocate starting with the smallest number of states appropriate to the application (e.g. one dry state and one wet state), to be increased only in response to model checking inadequacy. The aim is to choose the smallest number of clone states with which the observed dry-state persistence is captured well, and also the smallest number of wet states with which the marginal rainfall distribution is captured well.

3.2 Conditional rainfall model

Having specified a non-stationary (non-homogeneous) and essentially semi-Markovian latent structure, it remains to define a conditional model for rainfall occurrence and intensity. First recall the notation z_t, the latent state at any given time point. Continuing with our three state example, z_t takes only 3 values (dry, wet, wetter), noting that the dry state is made

up of D clone dry states, among which the conditional model is the same.

As zero rainfall is generally a common observation (approximately 88% of all observations in the Exeter time series), it makes sense to mix a continuous distribution for rainfall intensity with a probability mass at zero. This probability of zero rainfall, π_t, should vary with the latent state z_t – for example it should be higher in the dry state than in the wet states – and may also vary with time and/or climatological covariates. We achieve this by employing a logistic model for π_t:

$$\log\left(\frac{\pi_t}{1-\pi_t}\right) = v(t, z_t). \tag{8}$$

For our application to the Exeter gauge, we once again employ a combination of random effects and smoothing splines:

$$\begin{aligned} v(t, z_t = \text{dry}) &= \eta(\text{dry}) + b_1(t) + b_2(t) + m_b(t) + y_b(t); & (9)\\ v(t, z_t = \text{wet}) &= \eta(\text{wet}); & (10)\\ v(t, z_t = \text{wetter}) &= \eta(\text{wetter}), & (11) \end{aligned}$$

where in the same vain as (7) the dry state zero-probability is a sum of a smooth time-of-day spline $b_1(t)$, a smooth time-of-year spline $b_2(t)$, monthly i.i.d. Gaussian random effects $m_b(t)$ with variance $\delta^2_{m_b}$, and yearly random effects $y_b(t)$ with variance $\delta^2_{y_b}$. Parameter $\eta(z_t)$ is an intercept term, noting that the two wet states do not have any temporal structure in the model for the zero probability. This modelling choice was made after including equivalent temporal structures in (10)–(11) as in (9), but concluding these were not capturing anything on the basis of credible interval widths.

As discussed previously, there are many choices for the distribution of rainfall intensity, including Gamma, Weibull, Log-Normal and hybrid distributions. One of the key advantages of the approach we advocate is that it is possible to choose from any of these or other distributions, even using different distributions for each state if desired. Recalling that one of our key modelling aims is to capture extreme values well, we opt for a zero-location (zero-threshold) Generalized Pareto distribution (GPD) model, with scale parameter σ_t:

$$\begin{aligned} \log\Big(\sigma_t(z_t = \text{dry})\Big) &= \alpha(\text{dry}); & (12)\\ \log\Big(\sigma_t(z_t = \text{wet})\Big) &= \alpha(\text{wet}) + c_1(t, \text{wet}) + c_2(t, \text{wet}) + m_c(t, \text{wet}) + y_c(t, \text{wet}); & (13)\\ \log\Big(\sigma_t(z_t = \text{wetter})\Big) &= \alpha(\text{wetter}) + c_1(t, \text{wetter}) + c_2(t, \text{wetter}) + m_c(t, \text{wetter}) & (14)\\ &+ y_c(t, \text{wetter}); \end{aligned}$$

and with shape parameter ξ_t:

$$\begin{aligned}
\xi_t(z_t = \text{dry}) &= \gamma(\text{dry}); && (15)\\
\xi_t(z_t = \text{wet}) &= \gamma(\text{wet}) + g_1(t, \text{wet}) + g_2(t, \text{wet}) + m_g(t, \text{wet}) + y_g(t, \text{wet}); && (16)\\
\xi_t(z_t = \text{wetter}) &= \gamma(\text{wetter}) + g_1(t, \text{wetter}) + g_2(t, \text{wetter}) + m_g(t, \text{wetter}) && (17)\\
&+ y_g(t, \text{wetter}).
\end{aligned}$$

Once more we make use of time-of-day ($c_1(t, z_t)$ and $g_1(t, z_t)$) and time-of-year ($c_2(t, z_t)$ and $g_2(t, z_t)$) smooth splines, alongside monthly ($m_c(t, z_t)$ and $m_g(t, z_t)$) and yearly ($y_c(t, z_t)$ and $y_g(t, z_t)$) random effects, to capture inhomogeneity. Note that no temporal structure is assumed for the shape and scale parameters in the dry state, again a choice made in response to model checking.

The inclusion of independent splines and random effects for each state in all of the parameters of the conditional model (in this case π_t, σ_t and ξ_t) offers a high degree of flexibility for capturing diurnal, seasonal and temporal variation in the rainfall distribution. Moreover, capturing between-year variability with random effects (both here and in the dry state persistence) presents the opportunity to simulate effects for 'new' years from $\text{N}(0, \delta_y^2)$. This allows for a broader understanding of how rainfall, and any hydrological consequences, may vary in 'mild' and 'severe' years, compared to simulating new years with identical annual properties to the data.

In our dataset, the hourly observations are rounded to the nearest 0.2mm, which means that the likelihood should be adjusted accordingly. For example, if a rainfall observation is 2mm, the contribution to the likelihood should not just be the GPD density $f(2\text{mm}; ...)$, but should instead be $P(1.9\text{mm} < X \leq 2.1\text{mm} \mid ...)$. Furthermore, we truncate the GPD at 0.1mm such that the zero probability accounts for all values less than 0.1mm (which would be rounded to zero), such that the complete density function is:

$$\begin{aligned}
f(x; \pi, \sigma, \xi) &= \begin{cases} \pi & x = 0 \\ (1-\pi)\frac{F(x+0.1)-F(x-0.1)}{1-F(0.1)} & x = 0.2, 0.4, ... \end{cases} && (18)\\
F(x; \sigma, \xi) &= 1 - \left(1 + \frac{\xi x}{\sigma}\right)^{-\frac{1}{\xi}}, && (19)
\end{aligned}$$

where F is the cumulative distribution function of the zero-location GPD.

3.3 Prior distributions and implementation

We apply the model to hourly observations from the Exeter International Airport rainfall gauge, a time series of 70128 values spanning the 8-year period 2010 to 2017. To keep the model as general as possible, we specified uniform Dirichlet($\mathbf{1}$) prior distributions for transition matrix parameters $\boldsymbol{q}$, $\boldsymbol{v}$, $\boldsymbol{r}_1 = \{r_{1,0}, r_{1,1}, r_{1,2}\}$ and $\boldsymbol{r}_2$. We also specified non-informative Normal$(0, 10^2)$ prior distributions for the intercept parameters $\iota(d)$, $\eta(z_t)$, $\alpha(z_t)$, and $\gamma(z_t)$, where we use $d = 1, 2, 3$ clone dry states. However, a common problem with hidden Markov models is label switching, where the conditional models of one or more states swap. When this happens, the overall model is the same but parameter inference is convoluted, especially in a Bayesian implementation where Markov Chain Monte Carlo (MCMC) is employed. To prevent this, we impose the following constraints on the intercept parameters:

$$\iota(1) > \iota(2) > \iota(3); \tag{20}$$

$$\eta(\text{dry}) > max(\eta(\text{wet}), \eta(\text{wetter})); \tag{21}$$

$$\sigma_0(\text{wetter})(2^{\xi_0(\text{wetter})} - 1)\xi_0(\text{wetter})^{-1} > \sigma_0(\text{wet})(2^{\xi_0(\text{wet})} - 1)\xi_0(\text{wet})^{-1}. \tag{22}$$

These constraints relate to the intercepts, i.e. constrain what happens on average (since all temporal structures are centered at zero). They do not really restrict the model, they simply a) order the clone dry states (20), b) specify that the dry state will on average have higher probability of zero rainfall than the two wet states (21), and c) ensure the overall median rainfall implied by the intercepts (i.e. $\sigma_0 = \exp(\alpha)$ and $\xi_0 = \gamma$) is higher in the wetter state than in the wet state (22).

All splines were set up using the `jagam` function in the `mgcv` package for the programming language R (R Core Team, 2019). We specified 8 equidistant knots for the time-of-day splines and 6 for the time-of-year splines. For each spline, the coefficients are assigned Multivariate-Normal priors (Wood, 2016) where the covariance matrix is scaled by a parameter ν^2 (unique for each spline), which acts as a smoothing penalty (where smaller values of ν^2 correspond to a stricter penalty or more smoothing). More generally, this penalty is intended to avoid over-fitting, but here we would like spline effects to be quite smooth, leaving any non-smooth structured variability to be captured by the monthly and yearly random effects, and any short to medium term variability to be captured by the HMM latent state z_t. Thus for each ν we specified a Half-Normal$(0, \sqrt{2^2})$ prior, which corresponds to a modest smoothness penalty. The same Half-Normal priors were assigned to the random effect standard deviation parameters (δ_{m_b} and so on).

The model was implemented using the `nimble` package (de Valpine et al., 2017), a comprehensive suite for flexible MCMC inference. In this case, we needed to create a custom likelihood function that incorporates a version of the recursive forward algorithm used to compute the marginal likelihood in HMMs (Scott, 2002), thus avoiding sampling the latent states. The algorithm was adapted to allow for a temporally-varying transition matrix. We ran four MCMC chains in parallel, each for a total of 20k iterations, and discarded the first 10k as burn-in. After thinning by 10, we obtain 4000 samples in total. Owing to the complexity and size of the model (70128 hours of data), the model takes just over 12 hours to run on a high-end desktop in 2019. Each chain was randomly initialised and was assigned a different random number generator seed. Convergence of the four chains was assessed by visual inspection of trace plots and by computing the Potential Scale Reduction Factor (MPSRF) (Brooks and Gelman, 1998) for the following parameters: the initial state probability P_0; static transition matrix parameters ($\boldsymbol{q}$, $\boldsymbol{v}$, $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$), all the intercepts ($\boldsymbol{\iota}$, $\boldsymbol{\eta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$); and all the spline coefficients and random effects. This metric compares the within-chain and between-chain variance. Similar variance values typically result in a PSRF close to 1. Starting from different initial values and obtaining a PSRF close to 1 (less than 1.05 by convention) is a good indication of convergence. Here, the median PSRF across this set of parameters was 1.00, with a mean of 1.01, suggesting the chains have converged. All of the code to prepare and run the model is provided as supplementary material.

4 Model Checking and Results

In this section we assess model performance through comprehensive model checking and analysis of some key results. We simulate a new time series of length 70128, $\tilde{\boldsymbol{x}}$, from each posterior sample to obtain 4000 simulated time series. The general principle of model checking then relies on assessing whether certain characteristics of the observed values $\boldsymbol{x}$ are extreme relative to simulations from the model (Gelman et al., 2014).

4.1 Persistence and temporal dependence

We begin by assessing long dry periods, one of the three key characteristics of rainfall identified as important in Section 1. We define dry periods in two ways: 1) periods of consecutive zero rainfall values and 2) periods where rainfall does not exceed 0.2mm in any hour. Then, for each simulated time series $\tilde{\boldsymbol{x}}$, we calculate the quantiles (in increments of 0.5%) of the dry period durations. Figure 2 shows the median predicted dry period quantiles, with 95% prediction intervals, compared to the observed quantiles for both dry-period definitions. The

model generally captures the dry-period duration distributions very well, with the median values tracking the observed values (diagonal line) closely up until the last few quantiles, which are still contained within the 95% prediction intervals.

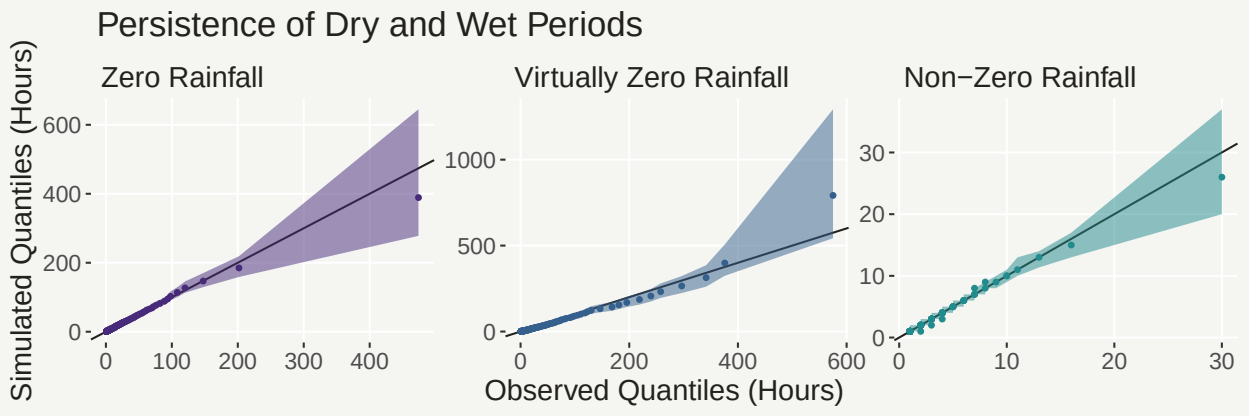

Figure 2: Median simulated quantiles of the durations of periods of zero rainfall (left), periods where rainfall does not exceed 0.2mm in any one hour (centre) and periods of non-zero rainfall (right), compared to observed duration quantiles for each, and with associated 95% prediction intervals.

In developing our approach, we found that the inclusion of 'clone' dry states made a dramatic improvement over the baseline HMM in capturing these distributions well – in fact the baseline was not able to capture any part of these distributions remotely well. However, it wasn't quite good enough until we allowed the dry persistence probabilities to vary with time. Recall that we did not include any clone wet states or non-stationarity in the 'wet' part of the transition matrix, so it is also important to check that the persistence of wet periods (defined as periods of consecutive non-zero rainfall values) is captured well. The median predicted quantiles of wet period durations are also shown in Figure 2 (right), and once more the whole distribution is captured very well.

We move on to assessing the temporal dependency structure, by considering the joint probability of rainfall exceeding some level l (mm) at two different time points separated by a lag of $k \geq 1$ hours, i.e. $\omega_{l,k} = P(X_t > l, X_{t+k} > l)$. We estimate this quantity for the observed data as the associated empirical frequency termed $\hat{\omega}_{l,k}$ and also for each of the $j = 1, \ldots, 4000$ simulated rainfall time series, termed $\tilde{\omega}_{l,k}^{(j)}$. For each l and k, $\hat{\omega}_{l,k}$ is compared to the posterior predictive distribution constructed from $\tilde{\omega}_{l,k}^{(j)}$. To that end, the proportion of simulated values less than or equal to the observed value, $h_{l,k} = (1/4000) \sum_{j=1}^{4000} \mathbb{1}(\tilde{\omega}_{l,k}^{(j)} \leq \hat{\omega}_{l,k})$, is calculated and then converted to a tail area probability by taking $1 - h_{l,k}$ if $h_{l,k} > 0.5$. As such, values of $h_{l,k}$ close to zero indicate $\omega_{l,k}$ is not well-captured. E.g., $h_{l,k} < 2.5\%$ means

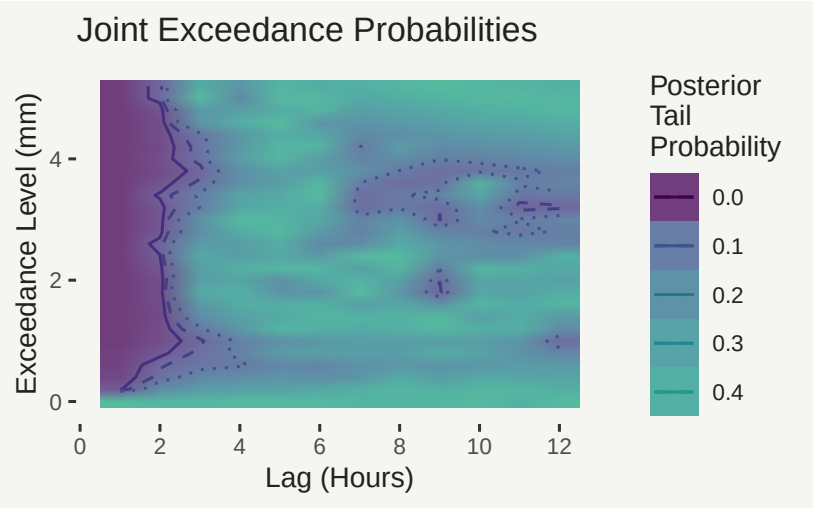

Figure 3: Posterior tail probabilities of observed joint empirical exceedance probabilities (as defined in the main text) compared to joint probabilities from 4000 simulated 8-year time series. The solid, dashed and dotted lines represent 2.5%, 5% and 10% contours of the tail probabilities, respectively.

that $\hat{\omega}_{l,k}$ is outside the 95% prediction interval. Figure 3 shows the tail area probabilities for $l = 0.2, 0.4, \ldots, 5.2$mm, where 5.2mm is the 99% quantile of the observed non-zero rainfall value, and for $k = 1, 2, \ldots, 12$ hours. Different colours/shades represent different $h_{l,k}$ values, while solid, dashed and dotted lines are contours at 2.5%, 5% and 10%, respectively. Generally, we would be concerned by values systematically less than 2.5% (the solid line), while values greater than 10% (the dotted line) are well within the bulk (80% prediction interval) of the predictive distribution. The leftmost region of the plot indicates a large area, encompassing one hour and two hour lags, where $h_{l,k}$ is close to zero. Specifically, the simulated joint probabilities in this region are too low, suggesting short-term temporal dependence is too weak. Elsewhere, however, $h_{l,k}$ is higher (with the exception of a few non-systematic patches, where the simulated joint probabilities are too high), meaning that apart from the very short term (1-2 hours), the model is generally able to capture the temporal dependency structure of the rainfall data well. We discuss one consequence of not capturing short-term dependence and how it might be improved in Section 5.

4.2 Seasonal and annual distributions

Whilst capturing the whole distribution of hourly rainfall values well, including extremes, can be challenging in itself, an even greater challenge is capturing this distribution as it varies by season. For example, many models overestimate extremes in the winter and underestimate them in the summer (Chandler et al., 2014).

First we check that the model is able to capture seasonal variation in the occurrence of rainfall. Figure 4 shows density plots of the proportion of zero values in each calendar season from the simulated time series $\tilde{\boldsymbol{x}}$, compared to the proportions in the observed values. All of

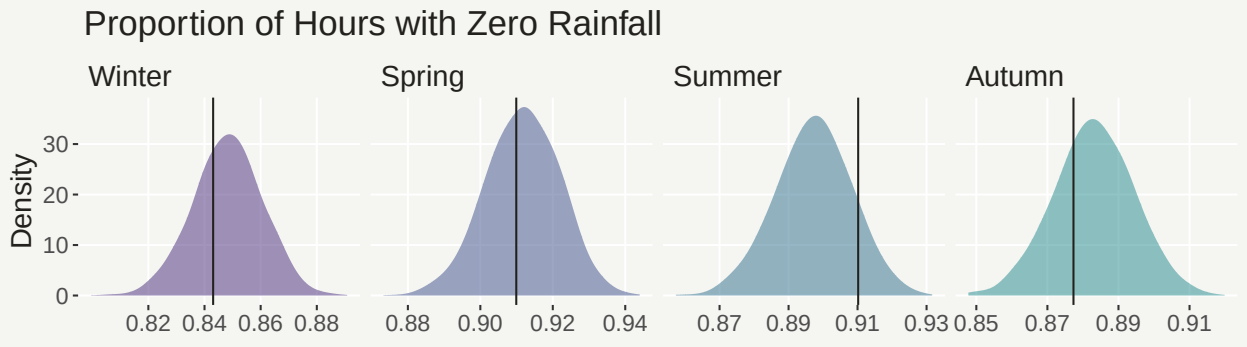

Figure 4: Density plots of the proportion of hours with zero rainfall, from 4000 simulated 8-year time series, by calendar season. The vertical lines represent the observed proportions.

the observed values are captured quite well. For example, the model captures the increased proportion of zeros in the summer compared to the winter, so it's clear that the model is able to reproduce seasonal variation in rainfall occurrence.

Next we check whether the model is able to capture seasonal variation in rainfall intensity. We do this for each season, by looking at quantile-quantile (Q-Q) plots of the sample quantiles (x-axis) and predicted quantiles (y-axis), defined as the median of the predictive distribution of each quantile. We also include the 95% prediction interval of each quantile and a 45° line. For adequate model fit, we would expect the points to be close to (but not exactly on) the line, and the line to be within the 95% prediction intervals. Figure 5 shows such a Q-Q plot for hourly rainfall values in each season. Clearly the distribution of rainfall values varies greatly by season, with higher (and more extreme) rainfall values in the summer and the autumn than in the winter and the spring. Despite this variation, the model is able to capture each season's distribution very well, all the way up to the extremes, especially given only two wet states were used. The model also performs generally well in capturing the distribution of hourly rainfall grouped by calendar month, owing to the inclusion of random effects, as illustrated by relevant Q-Q plots provided as supplementary material.

Lastly, Q-Q plots for rainfall values within each of the 8 years are shown in Figure 6. We note that all points are within the 95% prediction intervals and that there are a similar number of years where the few highest simulated quantiles are too high (2011, 2013) and years where the quantiles are too low (2010, 2014, 2016), suggesting the model is fitting individual years well, but not necessarily reproducing the data exactly. To check for over-fitting, one option would be to compare data from an out-of-sample year to simulations generated with new yearly effects y (drawn from their corresponding random effect distributions $\mathrm{N}(0, \delta_y^2)$ using Monte Carlo).

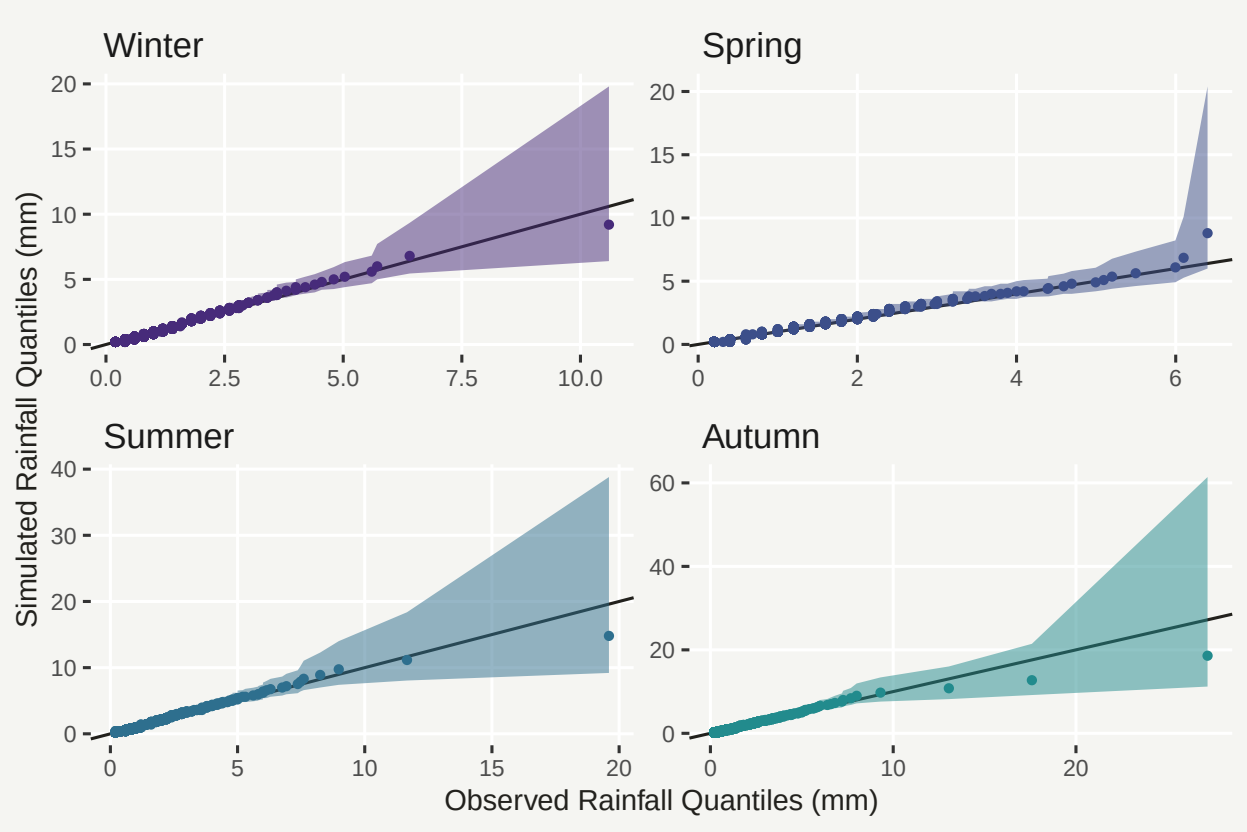

Figure 5: Median simulated versus observed non-zero rainfall quantiles (in increments of 0.1%), from 4000 simulated 8-year time series, by calendar season and with associated 95% prediction intervals.

4.3 Daily and monthly rainfall checking

For some applications it may also be important to check that the model captures the cumulative rainfall over a longer time period, such as a day, or a month. To assess this, we aggregate both the observed time series $\boldsymbol{x}$ and each simulated time series $\tilde{\boldsymbol{x}}$ (as detailed in Section 4) into daily and monthly totals. We can then assess whether the aggregated data ($\boldsymbol{x}_{daily}$ and $\boldsymbol{x}_{monthly}$) is extreme with respect to the simulated data ($\tilde{\boldsymbol{x}}_{daily}$ and $\tilde{\boldsymbol{x}}_{monthly}$).

First we check that the model captures the distribution of daily rainfall intensity well. The left panel of Figure 7 shows the corresponding Q-Q plot, indicating that the model captures the distribution very well, with surprisingly little deviation between the median predicted and observed daily values. The right panel of of Figure 7 shows the monthly Q-Q plot, which indicates that the model also captures the distribution of monthly values well, the only exception being the penultimate quantile, which we consider to be an acceptable deviation given that these are 95% prediction intervals. Indeed, both of these results are

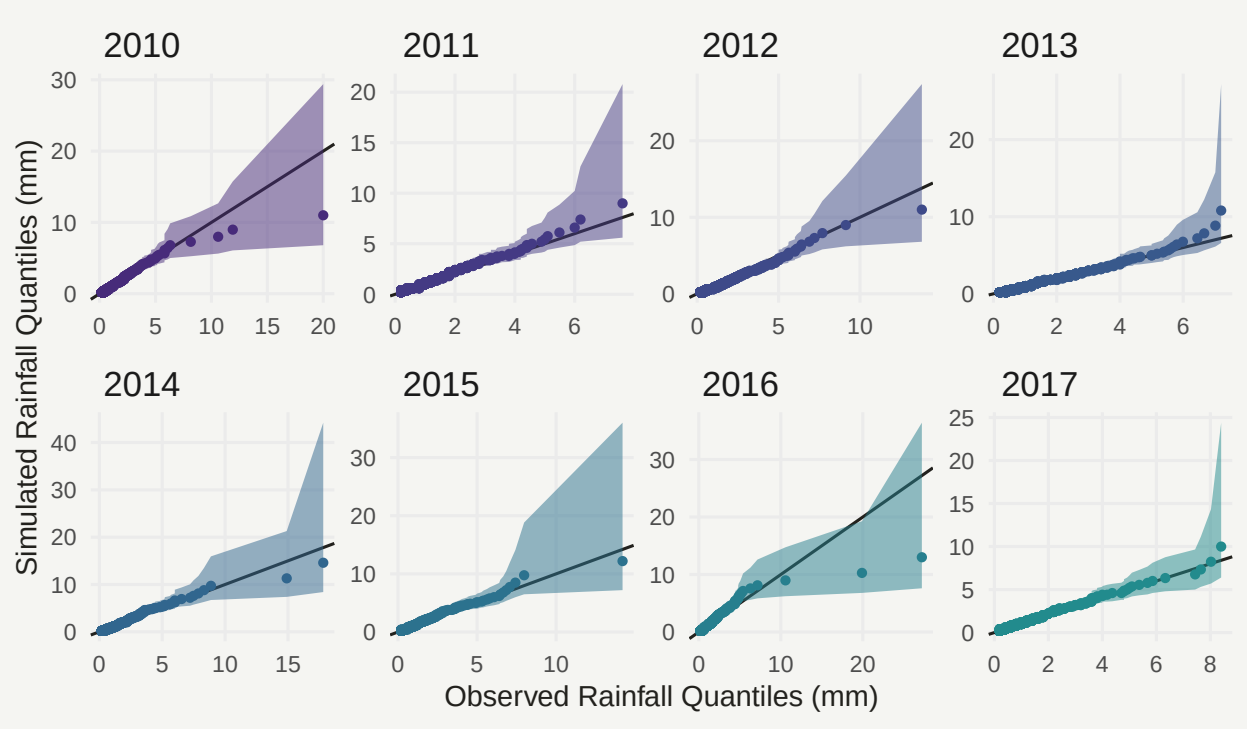

Figure 6: Median simulated versus observed non-zero rainfall quantiles (in increments of 0.1%), from 4000 simulated 8-year time series, by calendar year and with associated 95% prediction intervals.

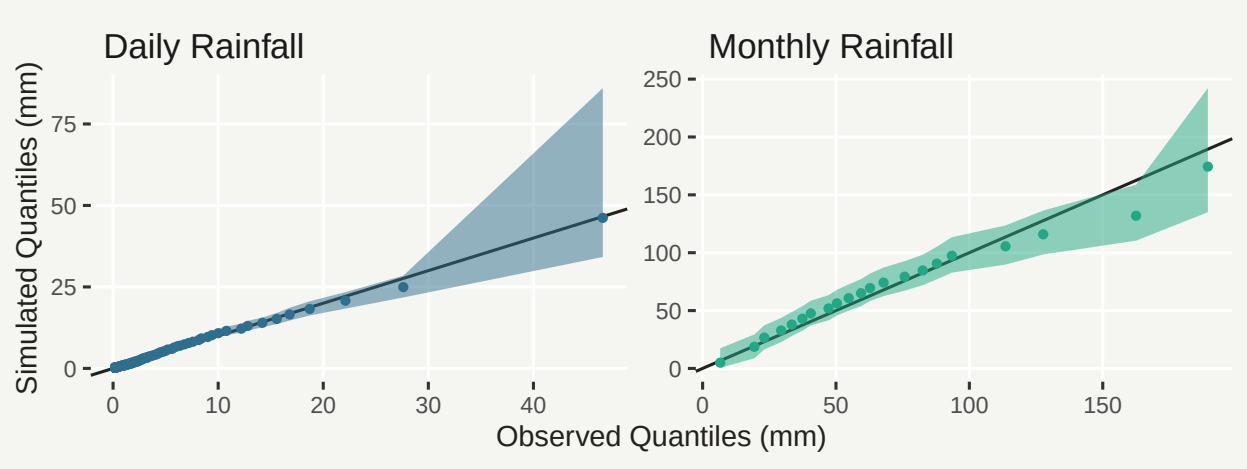

Figure 7: Median simulated daily (left) and monthly (right) versus observed non-zero rainfall quantiles (in increments of 1% for daily rainfall and 5% for monthly rainfall), with 95% prediction intervals.

impressive given that the model does not 'see' the daily or monthly totals, instead modelling them indirectly only through the hourly values.

4.4 Temporal effects

As well as aiding in fitting the data better, the temporal effects can indicate how different characteristics of rainfall vary with time-of-day, time-of-year, and between years. Figure 8 shows the posterior median predicted effects of time-of-day (left), time-of-year (middle) and year (right) on the expected persistence time of dry periods, as defined by (7). The time-of-day effect indicates a considerable dip in the persistence probability around midday, implying shorter dry periods in the afternoon. The time-of-year plot shows the smooth function $a_2(t)$ as a line and the sum of this and the monthly random effects $(a_2(t) + m_a(t))$ as a box plot. This indicates the expected behaviour – longer dry periods in the warmer months (May-September) – while not much monthly variability beyond what is explained by $a_2(t)$ (since boxes do not deviate much from the line). The yearly random effects show a fair amount of year-to-year variability with 2012 and 2014 exhibiting shorter dry period lengths compared to the other years. Notably, the time-of-year effect appears to dominate the others in terms of capturing variability in the dry-state persistence probability, indicated by its larger magnitude.

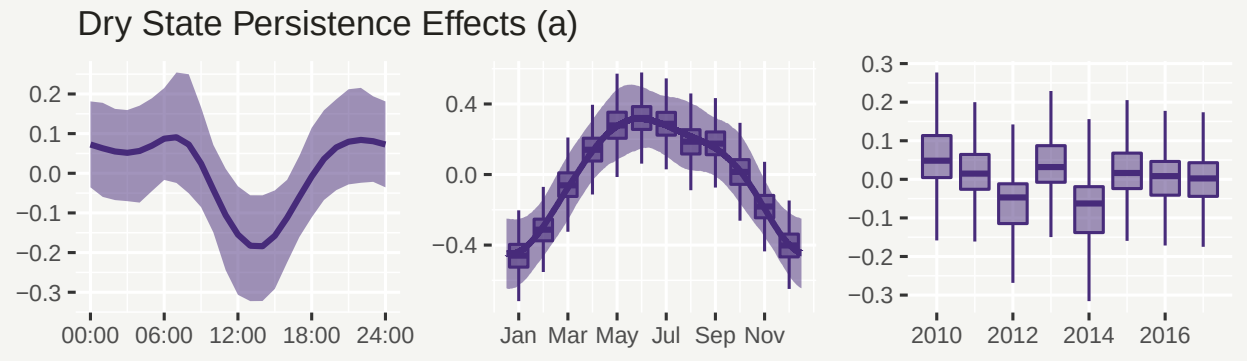

Figure 8: Posterior median predicted time-of-day (left), time-of-year (middle), and year-to-year (right) effects on the dry state persistence probability, along with 95% credible intervals.

Figure 9 shows the effects of time-of-day (left), time-of-year (middle) and year (right), on the conditional probability of zero rainfall (top row) and the distribution of rainfall intensity, through the Generalized Pareto scale and shape parameters (central and bottom rows, respectively) – see equations (12)–(17). We discuss these effects and their implications row-wise. **Dry state zero-probability (logit(π_t))**: Clearly there is a strong diurnal cycle

with a substantial drop in the zero-probability following midnight which slowly increases after midday. There is also a significant seasonal cycle, which as expected is highest during the summer months. Combined with the seasonal effect in Figure 8, the model suggests dry periods both last longer and have a higher probability of zero rainfall in the summer than in the winter. The yearly random effects indicate substantial but non-systematic variability across the 8 years. **Scale parameters ($\log(\sigma_t)$)**: The plots demonstrate strong but different diurnal cycles for both wet states, in addition to a fairly substantial seasonal cycle for the wetter state. While for the wet state the smooth seasonal spline is weak, the monthly random effects show noteworthy variability. Some year-to-year variability is also evident in both wet states. **Shape parameters (ξ_t)**: There is a strong diurnal cycle for the wetter state, which in conjunction with the scale parameter diurnal cycle implies lower median rainfall values in the afternoon but with a longer tail. A drop can be seen in the smooth seasonal cycles for both wet states in February and March, with substantial monthly deviation from the smooth line for the wetter state. Significant year-to-year variability is also evident, particularly for the wetter state.

4.5 Marginal rainfall distribution

Analysis of the spline and random effects in the previous subsection offers a clear 'under the hood' view of how the model captures different aspects of temporal variation. However, a better understanding of how simulations from the model behave over time can be obtained by analysing the marginal distribution of rainfall. For example, the GPD scale and shape effects work together to induce temporal variation in rainfall intensity, which can be summarised by investigating the marginal distribution of non-zero rainfall simulations.

Figure 10 shows how exceedance probabilities (conditional on the occurrence of rainfall) for hourly values simulated from the model ($\tilde{\boldsymbol{x}}$) vary with time of year. Notably, the median hourly value (exceedance probability of 0.5) does not change much over the coarse of the year, but the upper tail of the distribution is far heavier during the summer-autumn period than during the winter-spring period. In particular, the shape of the 0.05 exceedance probability line matches the seasonal splines for the scale and shape parameter in the wetter state, highlighting the link between the temporal effects and the marginal distribution of simulated rainfall.

Similarly, the zero-probability and dry state persistence effects work together to induce temporal variation in the occurrence of rainfall. Figure 10 also illustrates how the probability of rainfall in the hourly simulations varies with time of year. Here, the occurrence of rainfall is much less likely in the summer than in the winter. Notably, the shape of the curve is

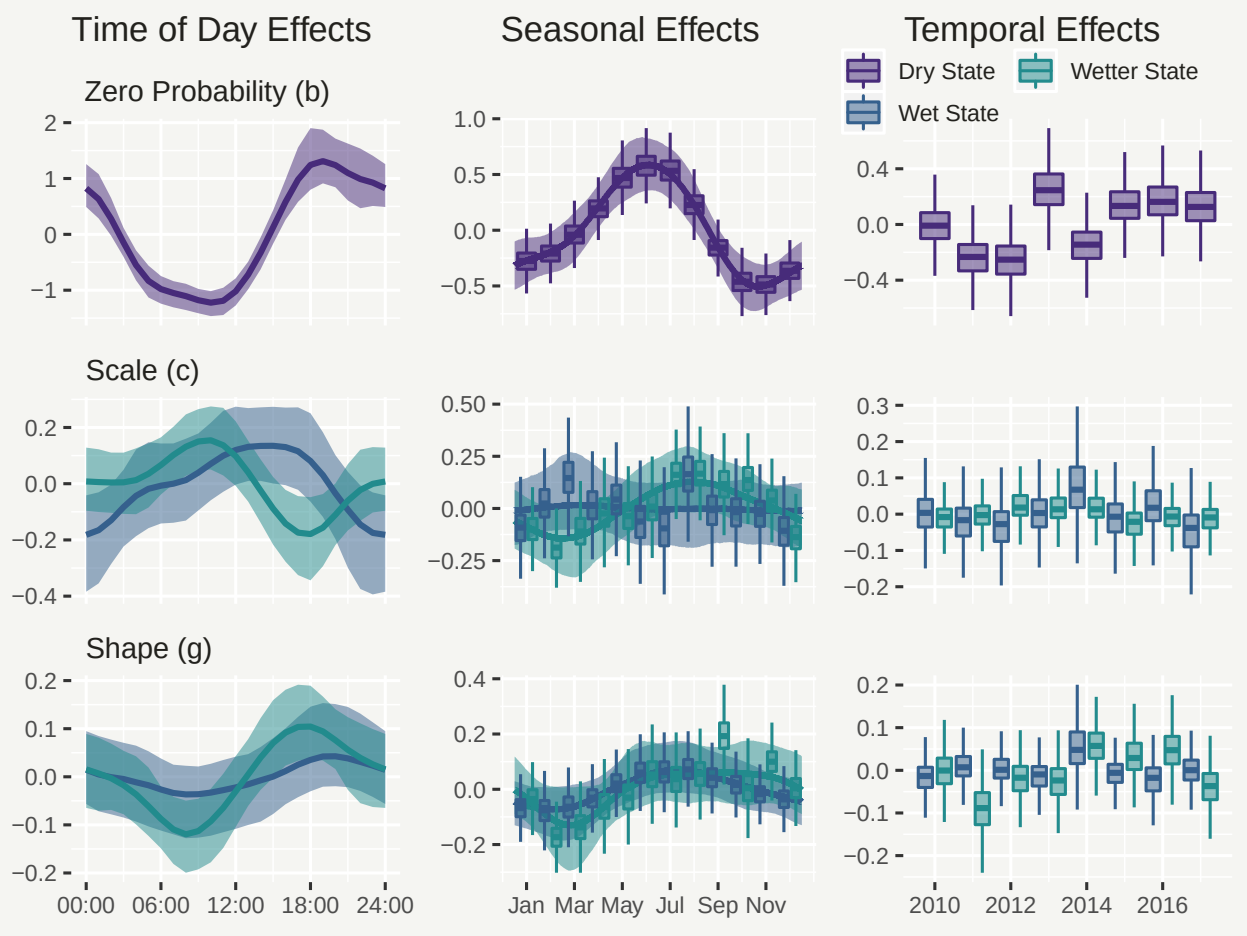

Figure 9: Posterior median predicted effects of time-of-day (left column), time-of-year (middle column) and year-to-year (right column) on the conditional probability of zero rainfall (top row), and the Generalized Pareto scale and shape parameters (central and bottom rows, respectively), by hidden Markov state and with 95% credible intervals.

more similar to the time-of-year/monthly effect in the dry persistence probability, including the kink in August, than the effect in the zero probability, suggesting the persistence of dry periods may play a greater role in the occurrence of rainfall. A plot equivalent to Figure 10 but for time-of-day is provided as supplementary material.

5 Discussion

In this article we discussed the role of stochastic rainfall modelling in the context of hydrological applications, such as urban flood modelling. We illustrated how the flexibility of the hidden Markov model framework allowed us to construct a compelling model for sub-daily

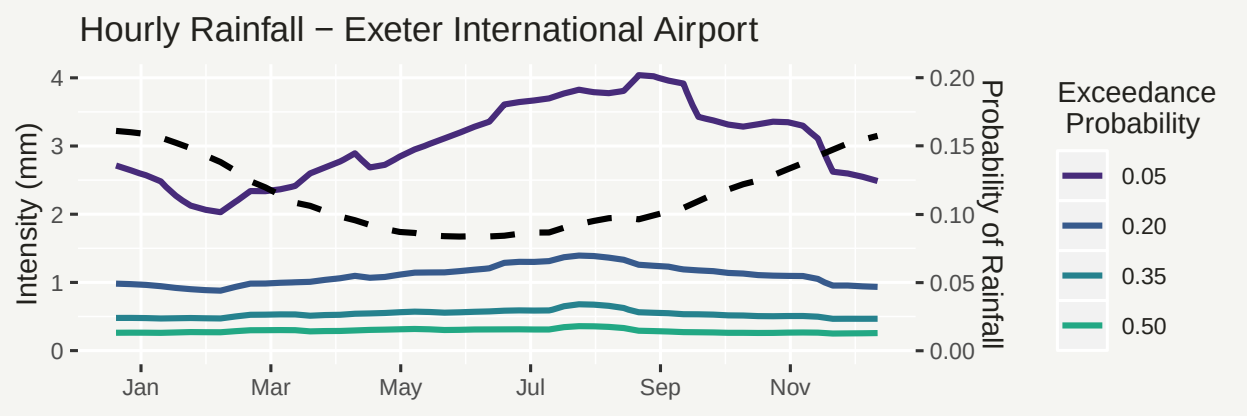

Figure 10: Posterior predictive exceedence probabilities of hourly rainfall (solid lines, left axis) and posterior predictive probability of rainfall occurrence (dashed line, right axis).

rainfall, which is able to capture all of the following crucial features: seasonal variation in rainfall occurrence and intensity; long dry periods; and extreme values. Our model incorporates several innovations compared to conventional approaches. These include clone states and temporal non-homogeneity in the transition matrix, which together allow the model to capture even the longest dry periods. Set in the Bayesian framework, our model also allows for a rich quantification of parametric and predictive uncertainty, meaning we can use posterior predictive checking to verify our model captures important characteristics of the data. In addition, the application to hourly rainfall data illustrates the applicability of the model in situations with high (temporal) resolution, something that is noticeably absent from the literature on direct rainfall models.

To demonstrate the effectiveness of our approach, we applied a relatively simple model comprising 3 clone dry states and 2 wet states to an 8-year long time series of hourly values from a rainfall gauge in Exeter, UK. We found that the model is able to capture well the distribution of dry period lengths, seasonal and annual variation in occurrence and intensity (including extreme values) and the distribution of intensity when aggregated to daily and monthly resolutions. We also illustrated how the model output can be interpreted in terms of how the rainfall occurrence and intensity change over the course of the day and year.

Although results were compelling, there remains some room for improvement. Most notably, simulations from the model displayed insufficient short-term (1-2 hours) temporal dependence. One consequence of this is that when aggregating hourly values into 3-hourly values, the model was able to capture the bulk of the distribution well but simulations in the upper-tail were systematically too low (as illustrated in Q-Q plots provided as supplementary material). This can be potentially improved by simply including more rainfall states,

although we suggest avoiding that in the interest of model interpretability. Alternatively, the model could be extended to incorporate an explicit dependency structure in the conditional models or indeed using "post-processing" ideas such as Gaussian anamorphosis. In addition, the model could be improved to better capture the most extreme values in some years (e.g. 2010, 2016) by utilising distributions that are more flexible compared to the GPD. For instance, Philippe et al. (2016), Chowdhury et al. (2017), or Tencaliec et al. (2018) are all approaches that aim to capture the entire range of rainfall (the bulk and extremes at either end).

We opted to apply the model to the particular time series from Exeter as it is situated in a region where floods pose a real risk to society, and because of the presence of several extreme values (arising from severe storms) which make modelling challenging. The inclusion of splines and random effects in potentially every part of the model affords a high degree of flexibility, in the sense that the states can change completely for different times of day and year, but also between years. That said, it is possible that there are some climates where the specific model we used to illustrate the framework may not be sufficiently flexible. In this case, the advantage of our approach is that it is fairly trivial to add more wet states, use alternative conditional distributions or indeed combine different conditional distributions, or incorporate more complex temporal structures (e.g. an interaction between the time-of-day and time-of-year splines), to name just a few potential adaptations. However, as with many statistical endeavours, this comes at the cost of increased complexity, so a balance must be struck to find a model which performs well enough without being impractical.

To that end, for new applications we advocate starting with the simplest possible version of the model presented here, such as a two state (dry, wet) HMM with some zero-inflated conditional models (e.g. zero-inflated GPD). Posterior predictive model checking should then indicate aspects of the data that the model does not capture adequately well. These inadequacies can then be addressed by a number of targeted model expansions (e.g. temporal structures to capture non-homogeneity, clone states to better capture the persistence of dry and wet periods), so as to avoid adding unnecessary complexity. The version of the model presented in this paper resulted from of a number of such iterations, including the addition of random effects and structured diurnal variability at the revision stage.

Finally, in this article we have focused on modelling the time series from one spatial location. To cater for applications where simulations at more than one location are required, future research will involve combining the innovations presented in this article with methods such as coupled hidden Markov models (Pohle et al., 2018), so that dependence between multiple spatial locations can be captured.

Supplementary Material

All R code required to implement the model is provided as supplementary material, alongside a simulated 8-year time series. Hourly rainfall data from the Exeter International Airport gauge is available upon registration at
`http://dx.doi.org/10.5285/7aaa582fb00246b794dc85950f1be265`.

Acknowledgements

This work was funded in part by a Natural Environment Research Council GW4+ Doctoral Training Partnership studentship [NE/L002434/1]. The authors also thank Christopher Paciorek, Berkeley, University of California, for helping to make the associated code more efficient, and two anonymous reviewers for their valuable comments which strengthened the paper considerably.

References

Antonello, M. and R. Roberto (2012). A mixed non-homogeneous hidden Markov model for categorical data, with application to alcohol consumption. Statistics in Medicine 31(9), 871–886.

Aryal, N. R. and O. D. Jones (2017). Fitting the Bartlett-Lewis rainfall model using approximate bayesian computation. Proceedings of the 22nd International Congress on Modelling and Simulation.

Brooks, S. P. and A. Gelman (1998). General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics 7(4), 434–455.

Bruno, B., B. Antonella, and C. Q. Antonio (2008). Using a hidden Markov model to analyse extreme rainfall events in Central-East Sardinia. Environmetrics 19(7), 702–713.

Chandler, R., V. Isham, P. Northrop, H. Wheater, C. Onof, and N. Leith (2014). Uncertainty in Rainfall Inputs, Chapter Chapter 7, pp. 101–152. Imperial College Press.

Chowdhury, A. F. M. K., N. Lockart, G. Willgoose, G. Kuczera, A. S. Kiem, and N. Parana Manage (2017). Development and evaluation of a stochastic daily rainfall model with long-term variability. Hydrology and Earth System Sciences 21(12), 6541–6558.

de Valpine, P., D. Turek, C. J. Paciorek, C. Anderson-Bergman, D. T. Lang, and R. Bodik (2017). Programming with models: Writing statistical algorithms for general model structures with NIMBLE. Journal of Computational and Graphical Statistics 26(2), 403–413.

Dewar, M., C. Wiggins, and F. D. Wood (2012). Inference in hidden Markov models with explicit state duration distributions. IEEE Signal Processing Letters 19, 235–238.

Dobson, A. and A. Barnett (2018). An Introduction to Generalized Linear Models. Chapman & Hall/CRC Texts in Statistical Science. CRC Press.

Economou, T., T. C. Bailey, and Z. Kapelan (2014, Sep). MCMC implementation for Bayesian hidden semi-Markov models with illustrative applications. Statistics and Computing 24(5), 739–752.

Environment Agency (2016). The costs and impacts of the winter 2013 to 2014 floods: Summary.

Environment Agency (2018). The costs of the winter 2015 to 2016 floods: Summary.

Furrer, E. M. and R. W. Katz (2008). Improving the simulation of extreme precipitation events by stochastic weather generators. Water Resources Research 44(12), n/a–n/a. W12439.

Gelman, A., J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin (2014, November). Bayesian Data Analysis, Third Edition (Chapman and Hall/CRC Texts in Statistical Science) (Third ed.). London: Chapman and Hall/CRC.

Glasbey, C. A. and I. M. Nevison (1997). Rainfall modelling using a latent gaussian variable. In T. G. Gregoire, D. R. Brillinger, P. J. Diggle, E. Russek-Cohen, W. G. Warren, and R. D. Wolfinger (Eds.), Modelling Longitudinal and Spatially Correlated Data, New York, NY, pp. 233–242. Springer New York.

Kim, Y. and G. Lee (2017, Aug). Stochastic precipitation generator with hidden state covariates. Asia-Pacific Journal of Atmospheric Sciences 53(3), 353–359.

Li, C., V. P. Singh, and A. K. Mishra (2012). Simulation of the entire range of daily precipitation using a hybrid probability distribution. Water Resources Research 48(3). W03521.

Meligkotsidou, L. and P. Dellaportas (2011, Jul). Forecasting with non-homogeneous hidden Markov models. Statistics and Computing 21(3), 439–449.

Met Office (2019). MIDAS Open: UK hourly rainfall data, v201901. Centre for Environmental Data Analysis.

Philippe, N., H. Raphael, R. Pierre, and H. Alexis (2016). Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. Water Resources Research 52(4), 2753–2769.

Pohle, J., R. King, M. van der Schaar, and R. Langrock (2018). Coupled Markov-switching regression: inference and a case study using electronic health record data. Proceedings of the 33rd International Workshop on Statistical Modelling 1, 242–246.

R Core Team (2019). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Rayner, D., C. Achberger, and D. Chen (2016). A multi-state weather generator for daily precipitation for the Torne River basin, northern Sweden/western Finland. Advances in Climate Change Research 7(1), 70 – 81.

Rodriguez-Iturbe, I., D. R. Cox, and V. Isham (1987). Some models for rainfall based on stochastic point processes. Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 410(1839), 269–288.

Scott, S. (2002). Bayesian methods for hidden Markov models: recursive computing in the 21st century. Journal of the American Statistical Association 97, 337–351.

Segond, M.-L., H. S. Wheater, and C. Onof (2007). The significance of spatial rainfall representation for flood runoff estimation: A numerical evaluation based on the Lee catchment, UK. Journal of Hydrology 347(1), 116 – 131.

Serinaldi, F. and C. G. Kilsby (2012). A modular class of multisite monthly rainfall generators for water resource management and impact studies. Journal of Hydrology 464-465, 528 – 540.

Spezia, L. (2006). Bayesian analysis of non-homogeneous hidden Markov models. Journal of Statistical Computation and Simulation 76(8), 713–725.

Tencaliec, P., A.-C. Favre, P. Naveau, and C. Prieur (2018). Flexible semiparametric Generalized Pareto modeling of the entire range of rainfall amount. Environmetrics, 1–22.

Wilks, D. S. and R. L. Wilby (1999). The weather generation game: a review of stochastic weather models. Progress in Physical Geography: Earth and Environment 23(3), 329–357.

Wood, S. (2016). Just another gibbs additive modeler: Interfacing jags and mgcv. Journal of Statistical Software, Articles 75(7), 1–15.

Yang, C., R. E. Chandler, V. S. Isham, and H. S. Wheater (2005). Spatial-temporal rainfall simulation using generalized linear models. Water Resources Research 41(11).

Zucchini, W., I. MacDonald, and R. Langrock (2017, July). Hidden Markov Models for Time Series An Introduction Using R, Second Edition (Chapman and Hall/CRC Monographs on Statistics and Applied Probability (Second ed.). London: Chapman and Hall/CRC.